Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 21 (2011) 230–239

Procedia Social and Behavioral Sciences

International Conference: Spatial Thinking and Geographic Information Sciences 2011

A New Areal Interpolation Method Based on Spatial Statistics

Daisuke Murakami[a*], Morito Tsutsumi[a]

[a]*Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Japan*

Abstract

Differences in spatial units among spatial data often complicate analyses. Spatial unit convergence, called areal interpolation, is often applied to address this problem. Of the many proposed areal interpolation methods, few consider spatial autocorrelation, which is the general property of spatial data. In this paper, by employing a spatial process model, a new areal interpolation method that considers spatial autocorrelation is presented. First, we briefly survey previous areal interpolation techniques and demonstrate that the stochastic method is superior to the deterministic method in archiving accurate interpolations. Next, after a discussion on the spatial process model, a new areal interpolation method is suggested. In this method, both spatial autocorrelation and the volume preserving property, a property that should be considered in areal interpolation, are considered using a combination of a linear regression based areal interpolation method, and the spatial process model. Finally, a case study on the areal interpolation of a population is provided to demonstrate that the suggested method succeeds in improving the predictive accuracy. This case study indicates that the consideration of spatial autocorrelation is important for accurate areal interpolation.

Keywords: Areal interpolation;Spatial dependence;Spatial statistics;EM-algorithm;Pycnophylactic property;

1. Introduction

Spatial data are often aggregated into spatial units. For example, national census data are released after being aggregated based on municipalities, while land-use data are often aggregated into specific grids. Because many types of spatial units exist for aggregation, differences in aggregation units often make handling data difficult. Hence, transferring spatial data from one zonal system to another is useful solving this problem. Such a process is called “areal interpolation” (e.g. [1,2]). We assume that areal interpolation is the conversion of spatial data from “source units” into “target units,” as shown in Fig.1. We call converted variables, **y**, “objective variables”.

Areal interpolation has been widely applied. In Japan, for example, grid unit census datasets are created by employing areal interpolation based on census data given from each basic unit district. However, most of the applied methods are simple and do not use supplementary data to explain the

* Corresponding author. Tel.: +81-29-853-5572; fax: +81-29-853-5070.
E-mail address: muraka51@sk.tsukuba.ac.jp.

1877–0428 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Yasushi Asami
Open access under CC BY-NC-ND license. doi:10.1016/j.sbspro.2011.07.034

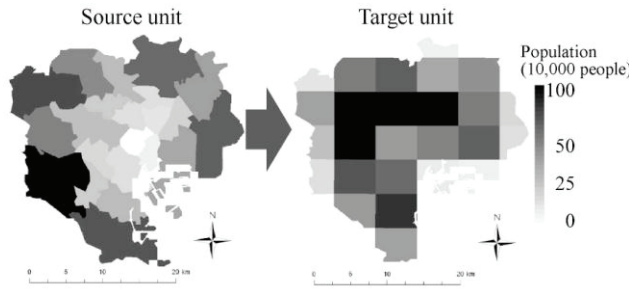


Fig.1. Outline of areal interpolation

distribution of objective variables. The accuracy of employing such interpolation methods is poor if the objective variables have a heterogeneous distribution [2].

On the other hand, areal interpolation methods that do use supplementary data have also been suggested. Representative examples of these methods are the dasymetric method [3], which utilizes land-use data, and the method of linear regression based method [4], which uses aggregated data. Because these methods explain the distribution of objective variables, they are more appropriate for accurate areal interpolation. Furthermore, the properties of spatial data should also be considered for accurate areal interpolation. Spatial dependence is a general property that implies that spatial data at nearby locations are similar, whereas those more widely separated are less similar (e.g., [5]). However, few studies have considered the spatial autocorrelation in areal interpolation (*e.g.* [6,7,8]).

This article proposes a new areal interpolation technique based on a spatial process model, a primal model in spatial statistics, which considers spatial autocorrelation. Then, this article examines the performance of the proposed method via a case study. Previous studies on areal interpolation are briefly surveyed in section 2. Section 3 explains the linear regression based areal interpolation method, which is an areal interpolation technique that serves as a basis of our method. Section 4 introduces the spatial process model, which is another important foundation of our method. The new areal interpolation method is described in section 5. Finally, in Section 6, the proposed method is applied to the areal interpolation of the population.

2. Areal Interpolation

Thus far, many studies have focused on spatial interpolation; however, most considered point data interpolation, which hereafter we call simply, "point interpolation." Point interpolation methods are generally conducted using Eq. (1):

$$y_s = f(s, \mathbf{x}_s), \tag{1}$$

where $s \in \boldsymbol{R}^2$ represents sites, $\mathbf{x}_s$ represents supplementary data, y_s is an objective variable, and $f(s, \mathbf{x}_s)$ is a function of s and $\mathbf{x}_s$. On the other hand, other studies have considered areal data interpolation based on areal data, called "areal interpolation." Areal interpolation methods are conducted using Eq. (2):

$$y_j = \int_{s \in B_j} f(s, \mathbf{x}_s) ds, \tag{2}$$

where B_j represents the j-th target unit and y_j is an objective variable in B_j.

When applying the point interpolation technique to areal data, the areal data are first replaced with point data. Next, $\hat{f}(s, \mathbf{x}_s)$ is identified based on the point data, and finally, interpolation is applied to fit $\hat{f}(s, \mathbf{x}_s)$ (Eq.1). For areal interpolation, on the other hand, objective variables observed in each source unit are first regarded as the aggregated value of the function of $f(s, \mathbf{x}_s)$, and the function is identified.

Next, y_j is interpolated by aggregating $\hat{f}(s, \mathbf{x}_s)$ into each target unit. Hence, interpolated values of each source unit, given by calculating back from the derived y_j, must equal to that real value. For example, an observed population in the source unit A_1 that comprises B_1 and B_2 must equal the sum of the population estimates in those two units. This property is called the "volume preserving property" (or pycnophylactic property: [1]) and is represented by Eq. (3):

$$\bar{y}_i = \int_{s \in A_i} f(s, \mathbf{x}_s) ds \qquad (3)$$

where A_i represents the i-th target unit and $\bar{y}_i$ is an objective variable in A_i.

Lam [9] indicated that the volume preserving property is the most basic property that should be satisfied in areal interpolation. Hence, we continue this discussion under the premise that areal interpolation is a data conversion technique from the source unit to the target unit through the application of Eq. (2), which satisfies the conditional equation of the volume preserving property, Eq.(3).

3. Linear-regression-based areal interpolation method

3.1. General procedure of areal interpolation

Most areal interpolation methods use spatial units whose boundaries are created using an intersection between the source and target units, and if necessary, units of additional data. We call these units "intersection units." Three units are presented as follows:

- Source unit: A_i ($i = 1,2,\ldots,I$)
- Target unit: B_j ($j = 1,2,\ldots,J$)
- Intersection unit: C_k ($k = 1,2,\ldots,K$)

Aggregated variables can be divided into two types: extensive and intensive variables. Extensive variable represents quantity whose value is proportional to the scale of spatial units such as population, while intensive variables represents quality whose value not depend on the scale such as population density. Because most spatial statistical models use intensity variables, we discuss areal interpolation of intensity variables hereafter. Objective variable y in each spatial unit is represented by vectors $\bar{\mathbf{y}}_i$, $\mathbf{y}_j$, and $\hat{\mathbf{y}}_k$ of sizes $I \times 1$, $J \times 1$, and $K \times 1$, respectively, where $I \le K$ and $J \le K$. Their elements are $\bar{y}_i$, y_j and $\hat{y}_k$, respectively.

$\hat{\mathbf{y}}_k$ is interpolated by two steps:

[1] The $\bar{\mathbf{y}}_i$ given in the source units are divided into intersection units and the $\hat{\mathbf{y}}_k$ given in the intersection units are interpolated.

[2] $\hat{\mathbf{y}}_k$ are aggregated into the target units.

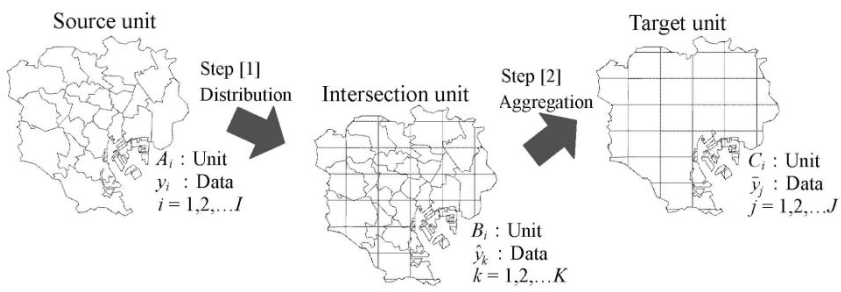


Fig.2. Procedure of areal interpolation

In the first, $\hat{y}_k$ must satisfy the constraint of the volume preserving property given in Eq. (4):

$$\bar{y}_i = \sum_{\{k|C_k \subset A_i\}} \frac{n_k}{n_i} \hat{y}_k \quad , \tag{4}$$

where n_i is the weight of A_i, and n_k is the weight of C_k. For example, if $\bar{y}_i$ represents the density variables divided by area, such as the population density, then n_i and n_k give the areas of each unit.

In contrast, in the second step, $\hat{y}_k$ must satisfy Eq. (9a):

$$y_j = \sum_{\{k|C_k \subset B_j\}} \frac{n_k}{n_j} \hat{y}_k \quad , \tag{5}$$

where n_j denotes the weight of the target unit B_j.

Because the value of y_j is decided automatically using Eq. (5), we will discuss the model employed in first step.

3.2. Linear regression based areal interpolation method

Flowerdew and Green (1992)[4] proposed a linear regression based areal interpolation method (we call LM based method hereafter) for density variables (e.g., population density). Their basic model is

$$\hat{y}_k = \sum_p x_{k,p}\beta_p + [\bar{y}_{i|A_i \supset C_k} - \sum_{k|C_k \subset A_i} \frac{n_k}{n_i} \sum_p x_{k,p}\beta_p] \quad , \tag{6}$$

where p denotes the index of explanatory variables, $x_{k,p}$ is the p-th explanatory variable given in the intersection unit k; and ε_k is the disturbance given in that unit. The first term of Eq. (6) represents the trend, and the second term represents the adjustment term in order to satisfy the volume preserving property. The second term is given based on Eq. (4).

The first step of areal interpolation, that is, data conversion between the source and the intersection units (see Seq.3.1.), is done by employing the EM algorithm [10], which allows maximum likelihood estimation using incomplete data. The calculation procedure is as shown below:

[1–1] Initial values of each $\hat{y}_k$ are set.

[2–2] The following steps are iterated until the parameters achieve convergence.

Maximization (M) step

The parameter β_p is estimated by applying weighted least squares as

$$\hat{y}_k = \sum_p x_{k,p}\beta_p + \varepsilon_k \qquad \varepsilon_k \sim N(0, \sigma^2 n_k) \, . \tag{7}$$

Expectation (E) step

By substituting the estimated β_p into Eq. (6), $\hat{y}_k$ is interpolated.

4. Spatial Process Model (SPM)

The spatial process model (SPM: [11]) is a well-known model in spatial statistics. The basic type of SPM is given through the following equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \, , \qquad \boldsymbol{\varepsilon} \sim N(0, \mathbf{C}) \tag{8}$$

where $\mathbf{y}$ is an $N \times 1$ vector of explained variables in space $\mathbf{s} \in R^2$ ($s = 1, \ldots, N$), $\mathbf{X}$ is an $N \times p$ matrix of the explanatory variables within the same sites, $\boldsymbol{\varepsilon}$ is an $N \times 1$ local variable vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the trend parameters and $\mathbf{C}$ is an $N \times N$ variance-covariance matrix. In order to consider spatial autocorrelation, elements of $\mathbf{C}$ are given by the function of distance called covariance function assuming that the elements of $\boldsymbol{\varepsilon}$, ε_s, are from a second-order stationary process. When isotropy holds, the process of ε_s satisfies the following equations

$$E[\varepsilon_s] = 0, \qquad (9) \qquad Cov[\varepsilon_s, \varepsilon_{s'}] = C(d_{s,s'}). \qquad (10)$$

where ε_s is the local variable of site s and $d_{s,s'}$ is the Euclidean distance between site s and site s'. There are three parameters that characterize the form of the covariance function: nugget (σ^2), partial– sill (τ^2), and range (w). Among the several proposed covariance functions, the exponential covariance function Eq.(11) is one of the most common.

$$C(d_{s,s'}) = \begin{cases} \tau^2 \exp\left[-\left(\dfrac{d_{s,s'}}{w}\right)\right] & (s \neq s') \\ \tau^2 + \sigma^2 & otherwise \end{cases} \qquad (11)$$

For more details about covariance function, see Cressie (1993)[5].

To predict the unknown value y_0 at arbitrary site $s_0 \in R^2$, let

$$y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \varepsilon_0, \qquad (12)$$

where y_0 is an explained variable, $\mathbf{x}_0$ is a p×1 matrix of the explanatory variables, and ε_0 is a local variable. Minimizing the expected square error yields the following equation:

$$\varepsilon_0 = \mathbf{c}_0'\mathbf{C}^{-1}\boldsymbol{\varepsilon}, \qquad (13)$$

where $\mathbf{c}_0$ is an N×1 covariance vector between observed site and predicted site. y_0 is given by Eq.(14) using Eqs.(8), (12) and (13):

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} + \mathbf{c}_0'\mathbf{C}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad where \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}. \qquad (14)$$

5. Areal Interpolation Method Based on SPM

5.1. Model construction

An areal interpolation method which considers both spatial dependence and volume preserving property is suggested by combining the basic model of LM based method SPM. Because SPM formulates spatially continuous process, we re-define intersection units subdividing the spatial units whose boundaries are created using an intersection between the source and target units into fine units so that we can regard them continuous. $\mathbf{y}_k$ given in each intersection unit are assumed to obey the isotropic stationary process and assume that $\mathbf{y}_k$ obey Eq.(15):

$$\hat{\mathbf{y}}_k = \mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k \qquad \boldsymbol{\varepsilon}_k \sim N(0, \mathbf{C}_k), \qquad (15)$$

where $\mathbf{X}_k$ is a K×p matrix of explanatory variables, and $\mathbf{C}_k$ is a K×K variance-covariance matrix. To assure the constraint of volume preserving property given by Eq.(7) under the condition of the spatial process Eq.(15), the following equation must be satisfied:

$$\overline{\mathbf{y}}_{i|A_i \supset C_k} = \mathbf{N}_k\mathbf{X}_k\boldsymbol{\beta} + \mathbf{N}_k\boldsymbol{\varepsilon}_k. \qquad (16)$$

where $\overline{\mathbf{y}}_{i|A_i \supset C_k}$ is the $K \times 1$ vector whose k'-th element is the objective variable of i'-th source units which contains k'-th intersection unit**.** $\mathbf{N}_k$ is a $K \times K$ block diagonal matrix. $\mathbf{N}_k$ has I blocks, and the i'-th block is the $K' \times K'$ matrix representing the weights of the k'-th intersection units, which are contained in the i'-th source unit. Each element of the k'-th column in the i'-th block is $n_{k'}/n_{i'}$. Imposing the conditional equation to satisfy the volume preserving property, Eq.(16), on Eq. (15), we have Eq. (17):

$$\hat{\mathbf{y}}_k = \mathbf{X}_k\boldsymbol{\beta} + [\overline{\mathbf{y}}_{i|A_i \supset C_k} - \mathbf{N}_k(\mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k)] + \boldsymbol{\varepsilon}_k \qquad \boldsymbol{\varepsilon}_k \sim N(0, \mathbf{C}_k), \qquad (17)$$

Eq. (17) is regarded as the model which extends the basic model of LM based method, Eq. (6), in order to consider spatial autocorrelation. Then, replacing Eq. (6) with Eq. (17) in calculation step [2] shown in Sec. 3.2., areal interpolation that considers both spatial autocorrelation and the volume preserving property is achieved.

5.2. Areal interpolation procedure employing the proposed model

We assume that an areal interpolation procedure employing our model obeys the same steps as the LM based method discussed in Sec. 3. In this section, an areal interpolation procedure based on our model for the interpolation step [1] (see Sec. 3.1.).

5.2.1. M‐step

To construct the likelihood function of our model, Eq. (17) is expanded and Eq. (18) is given:

$$(\mathbf{I}-\mathbf{N}_k)\boldsymbol{\varepsilon}_k = (\hat{\mathbf{y}}_k - \overline{\mathbf{y}}_{i|A_i \supset C_k}) - (\mathbf{I}-\mathbf{N}_k)\mathbf{X}_k\boldsymbol{\beta}\,. \tag{18}$$

It should be noted that $\mathbf{N}_k$ is a block diagonal matrix whose elements of each row that represent intersection units contained in the same source unit A_r are identical. Thus, $\mathbf{N}_k$ is generally not a fill rank matrix, and therefore matrix $(\mathbf{I}-\mathbf{N}_{ik})$ becomes singular. As a result, $\boldsymbol{\varepsilon}_k$ cannot be specified using Eq. (18) directory. A problem that cannot specify $\boldsymbol{\varepsilon}_k$ is an ill–posed problem. To solve this problem, we consider to apply the generalized inverse matrix (e.g. Aster et al. (2005)[12]); a quasi-inverse matrix given by optimizing the additional conditions. Thus, Eq. (18) is expanded into Eq. (19) employing the generalized inverse matrix of $(\mathbf{I}-\mathbf{N}_k)$, $(\mathbf{I}-\mathbf{N}_k)^+$:

$$\boldsymbol{\varepsilon}_k \approx (\mathbf{I}-\mathbf{N}_k)^+(\hat{\mathbf{y}}_k - \overline{\mathbf{y}}_{i|A_i \supset C_k}) - (\mathbf{I}-\mathbf{N}_k)^+(\mathbf{I}-\mathbf{N}_k)\mathbf{X}_k\boldsymbol{\beta}\,. \tag{19}$$

Optimization of the generalized inverse matrix is achieved by minimizing the residual sum of squares (as in Eq. (19), the sum of the squares of the differences of both sides of the equation), the norm (as in Eq. (23), $\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_k$), or another measure. Minimizing the residual sum of squares is important to assure the accuracy of the expansion of Eq. (18) into Eq. (19), whereas minimizing the norm is important to minimize the variance of $\boldsymbol{\varepsilon}_k$. The Moore–Penrose generalized inverse matrix given by Eq. (20) is a generalized inverse matrix for minimizing both the residual sum of squares and the norm.

$$\mathbf{M}^+ = \mathbf{M}'(\mathbf{M}'\mathbf{M}\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\,, \tag{20}$$

where $\mathbf{M}$ denotes a matrix containing singular and non‐square matrix. We expand Eq. (18) into Eq. (19) by employing Eq. (20).

The log–likelihood function of the traditional SPM is given by Eq. (21):

$$\ln(L) = l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X})'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,, \tag{21}$$

Intuitively, Eq. (19) is a special case of a traditional SPM Eq.(8), whose explained and explanatory variables are given by $\mathbf{a}_k=(\mathbf{I}-\mathbf{N}_{ik})^+(\hat{\mathbf{y}}_k - \overline{\mathbf{y}}_{i|A_i \supset C_k})$ and $\mathbf{B}_k=(\mathbf{I}-\mathbf{N}_{ik})^+(\mathbf{I}-\mathbf{N}_{ik})\mathbf{X}_k$ respectively. The log–likelihood function of our model is thus given by

$$l_k \approx -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}_k| - \frac{1}{2\sigma^2}(\mathbf{a}_k-\mathbf{B}_k\boldsymbol{\beta})'\mathbf{C}_k^{-1}(\mathbf{a}_k-\mathbf{B}_k\boldsymbol{\beta})\,. \tag{22}$$

Likelihood maximization is conducted in the same manner as a traditional SPM. That is, $\boldsymbol{\beta}$ and $\hat{\sigma}^2$ are first estimated using Eqs. (23) and (24), respectively. Then, $\hat{\tau}^2$ and $\hat{w}^2$ are given by maximizing their concentrated likelihood functions:

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}_k'\mathbf{C}_k^{-1}\mathbf{B}_k)^{-1}\mathbf{B}_k'\mathbf{C}_k^{-1}\mathbf{a}_k\,, \tag{23}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{a}_k-\mathbf{B}_k\hat{\boldsymbol{\beta}})'\mathbf{C}_k^{-1}(\mathbf{a}_k-\mathbf{B}_k\hat{\boldsymbol{\beta}})\,. \tag{24}$$

5.2.2. E-Step

Based on Eqs. (8), (14), and (17), the predictor of $\hat{\mathbf{y}}_k$ in Eq. (17) is given by the conditional expectation of $\hat{\mathbf{y}}_k$, $\hat{\mathbf{y}}_{k0}$:

$$E(\hat{\mathbf{y}}_{k0} \mid \hat{\mathbf{y}}_k) = \overline{\mathbf{y}}_{i|A_i \supset C_k} + (\mathbf{I} - \mathbf{N}_k)\{\mathbf{X}_k\hat{\boldsymbol{\beta}} + \mathbf{C}'_{0k}\mathbf{C}_k^{-1}(\mathbf{a}_k - \mathbf{B}_k\hat{\boldsymbol{\beta}})\}, \quad (25)$$

where $\mathbf{C}_{0k}$ is the cross covariance between ε_k and the local variables contained in the elements of $\hat{\mathbf{y}}_{k0}$: ε_{0k}. Here, we assume that ε_k and ε_{0k} obey the same spatial process explained by $\hat{\tau}^2$ and $\hat{w}^2$, and that ε_{0k} does not have a nugget, σ^2. This is because traditional SPM–based point interpolation also assumes that ε_i and ε_0 obey the same process, and that ε_0 does not have σ^2. Hence, these assumptions are natural. Based on their assumptions, the elements of $\mathbf{C}_{0k}$ are given by the covariance function without σ^2, e.g., Eq. (26) is an example based on the exponential covariance function:

$$C(d_{k,k'}) = \tau^2 \exp\left[-\left(\frac{d_{k,k'}}{w}\right)\right]. \quad (26)$$

5.2.3. Calculation procedure

Areal interpolation based on the constructed model is conducted using the EM-algorithm in the following manner:

[1] Estimate parameters $\boldsymbol{\beta}$, $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{w}^2$ by applying the traditional SPM whose explained variables are $\overline{\mathbf{y}}_{i|A_i \supset C_k}$ and explanatory variables are $\mathbf{N}_k\mathbf{X}_k$.

[2] The following steps are repeated until convergence:

E–Step

Update $\hat{\mathbf{y}}_k$ by the conditional expectation $\hat{\mathbf{y}}_k$ of given by Eq. (25).

M–Step

Maximize the log-likelihood function Eq. (22).

Applying these procedures, the predictor, $\hat{\mathbf{y}}_k$, which considers spatial autocorrelation and the volume preserving property, is given in each intersection unit. Then, the explained variables of the target units $\mathbf{y}_j$ are interpolated substituting $\hat{\mathbf{y}}_k$ in Eq. (5).

5.2.4. Comparison of areal interpolation methods considering spatial autocorrelation

Some areal interpolation methods that consider spatial autocorrelation have been recently proposed. They contain SPM-based methods (e.g. [6,13,14,7]); the Bayesian method, which considers a conditional autoregressive process [15]; and a method based on a spatial econometric model [8]. SPM-based methods have a merit in that their predictors minimize the expected square error, whereas the merit of the Bayesian method is to obtain the predictive distribution. Finally, the benefit of the method based on spatial econometrics is its adaptability to socio-econometric data. In SPM-based methods, Kiriakidis [6], Yoo and Kyriakidis [13,14], Gotway and Young [7] consider the volume preserving property by restricting the weight of local variables given in observed sites, Eq. (13), whereas our method restricts the basic equation of the traditional SPM, Eq. (8). As a result, former methods minimize the expected square error of $\hat{\mathbf{y}}_k$ itself, whereas our method minimizes that of $(\mathrm{I} - \mathrm{N}_{ik})^{+}(\hat{\mathbf{y}}_k - \overline{\mathbf{y}}_{i|A_i \supset C_k})$. Then, the influence of the difference in the criteria of minimization should be discussed in future research.

6. Case Study

6.1. Model and Dataset

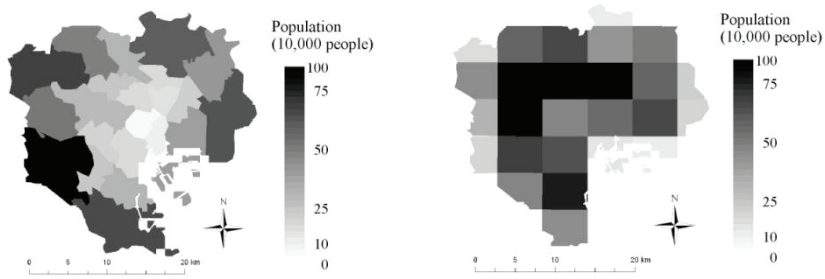


Fig. 3. Population in the source (23 wards of Tokyo: left) and target units (5km grids: right)

To examine the applicability of the proposed model, we apply it to census population (night time) for the 23 wards of Tokyo, Japan, from 2007, provided by the Ministry of Internal Affairs and Communications (MIC). Because population is extensive variable whose value depends on the scale of spatial units (see Seq. 3.1), we first convert it into population density, then, interpolate population density using the proposed method. Finally, we calculate back population from the estimated population density. In this study, the LM based method (Non-spatial), which considers the volume preserving property, and the method suggested in this paper (Spatial), which considers both the volume preserving property and spatial autocorrelation, are compared. They are applied to areal interpolation of a population whose source units and target units are the 23 wards of Tokyo and 5km grids respectively (Fig. 3). The intersection units are defined based on the intersection of the source unit, target unit, and 1 km grid. To check the accuracy of methods, the observed populations given in the 5km grids are required. Hence, we construct them to aggregate the census population (night time) given in 1 km grids provided by MIC into the 5km grids. For the explanatory variables, we assume the distance from the nearest station (Sta_dist) [km]; distance between the nearest station and the city center, more precisely, the four largest stations in terms of daily passengers (Tokyo, Shinjuku, Ikebukuro, and Shibuya stations) (City_dist) [km]; and the logarithm of the officially accessed land price (LP) [thousand yen per square meter] at the nearest site in a residential area. These data are constructed based on digital national land information provided by the Ministry of Land, Infrastructure, Transport, and Tourism. Among the several possible covariance functions, an exponential covariance function is applied in this study.

The calculations in this study were performed in R 2.11.1 (provided by CRAN), which is a free software for statistical computing. ArcGIS 9.3.1 (provided by ESRI) was used for spatial map creation.

6.2. Parameter estimation

Table 2 shows the results of the parameter estimates achieved by both models. The table shows that the estimates of the trend parameters for (Spatial) tend to take a larger absolute value compared with the estimates for (Non-spatial). The sign of the trend parameter of City_dist, which was expected to have a negative value, is positive for (Non-spatial), whereas it is negative for (Spatial). The (Spatial) estimate of the range indicates the existence of spatial autocorrelation with a range of 11.4 (=3.8×3) km.

Table 1 Result of parameter estimation

Parameter	Non-spatial	Spatial
Sta_dist (km)	26.9 (–2.31)	–148.752 (–12.2)
City_dist (km)	14.2 (1.07)	–54.183 (–4.09)
LP (thou. yen/m^2)	–209 (–2.04)	–826.369 (–8.10)

Parameter	Non-spatial	Spatial
Nugget		0.001
Partial–sill		0.001
Range		3.80

[1)] A value in parentheses represents t value

6.3. Parameter estimation

Table 2 shows the results of the parameter estimates achieved by both models. The table shows that the estimates of the trend parameters for (Spatial) tend to take a larger absolute value compared with the estimates for (Non-spatial). The sign of the trend parameter of City_dist, which was expected to have a negative value, is positive for (Non-spatial), whereas it is negative for (Spatial). The (Spatial) estimate of the range indicates the existence of spatial autocorrelation with a range of 11.4 (=3.8×3) km.

6.4. Test of predictive accuracy

The predictive accuracy is measured using the root mean square error (RMSE) based on the interpolated value $\hat{\mathbf{y}}_j$ given in 5-km grids as follows:

$$RMSE = \sqrt{\sum_j \frac{(y_j - \hat{y}_j)^2}{J}}, \tag{27}$$

where j represents the number of intersection units whose number of units is J. Table 3 shows both RMSE values. Here, the gap between two models represents the effect when considering spatial autocorrelation. Their results show the importance to consider spatial autocorrelation in areal interpolation.

Figs. 4 show plots for the interpolated values given in each target unit. The plots given in the 5km grids are similar with the observed values shown in Fig. 5. To quantitatively compare the predictive accuracies of (Non-spatial) and (Spatial), gaps in the error ratio (ER) are plotted in Fig. 5. The ER for each grid is given as follows:

$$ER = 100 \times \left| \frac{y_j - \hat{y}_j}{y_j} \right|. \tag{28}$$

Fig. 5 shows that (Spatial) is more accurate than (Non-spatial) especially in the north and west region, whereas it is less accurate in the mid and east area. Improvements in the north-western area may be caused by the homogeneity of this area, which is consistent with the assumption of the second-order stationarity assumed in (Spatial), whereas aggravation in the mid area may be caused by the heterogeneity of this area, which is not consistent with that assumption.

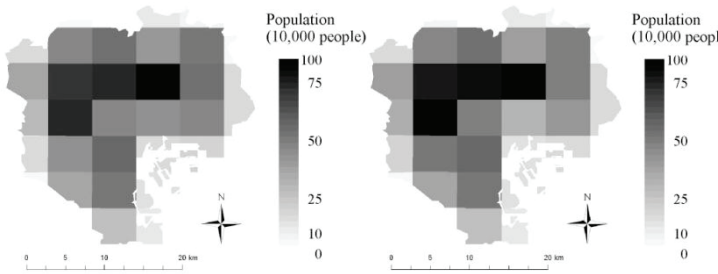
7. Conclution

We proposed a new areal interpolation method that considers the volume preserving property and spatial autocorrelation. This method is based on the SPM and LM based method, which is a primal statistical areal interpolation technique. A case study showed that the suggested method succeeds in improving the predictive accuracy, and indicated that a consideration of the spatial autocorrelation is important for an accurate areal interpolation.

On the other hand, as shown in the case study, the relationship between areal interpolation accuracy and the scale of the target units should be clear. Furthermore, the relationship between the accuracy and data properties should also be clear.

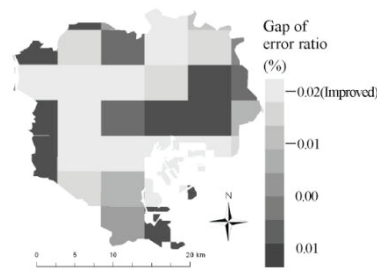
Table 2 Result of parameter estimation

	Non-spatial	Spatial
RMSE (5km grid)	103	95.2



Non-spatial Spatial

Fig.4. Interpolated population given in the 5km grids



(ER of Spatial) – (ER of Non-spatial)

Fig.5. Gap of the error ratio between methods

In a future study, we want to develop our method to consider temporal dependence in addition to spatial autocorrelation. Spatial heterogeneity, as shown in the mid area of Tokyo in the case study, is also an important property, one that is desirable for implementation in our model.

A comparison among other areal interpolation methods is also an important subject.

Acknowledgements

This study was supported by a Grant-in-Aid for Scientific Research (B) (23360219).

References

[1] Tobler W R. Smooth Pycnophylactic Interpolation for Geographical, Regions. *Journal of the American Statistical Association*; 1979; (**74**):519–530.

[2] Sadahiro Y. Accuracy of Areal Interpolation: A Comparison of Alternative Methods. *Journal of Geographical Systems*, 1999; **1** (4): 323–346.

[3] Wright J K. A Method of Mining Densities of Population with Cape Cod as an Example. *Geographical Review*; 1936; **26**:103–110.

[4] Flowerdew R, Green M. Developments in Areal Interpolation Methods and GIS. *Annals of Regional Science*, 1992; **26**: 67–78.

[5] Cressie N. *Statistics for Spatial Data*. Revised Edition, John Wiley & Sons; 1993.

[6] Kyriakidis P C. A Geostatistical Framework for Area-to-Point Spatial Interpolation. *Geographical Analysis*; 2004; **36**(3): 259–289.

[7] Gotway C A, Young L J. A Geostatistical Approach to Linking Geographically Aggregated Data From Different Sources. *Journal of Computational and Graphical Statistics*; 2007; **16**(1):115–135.

[8] Tsutsumi M, Murakami D. A New Areal Interpolation Technique Based on Spatial Econometrics. 4th World Conference of Spatial Econometrics Association, Chicago, USA; 2010.

[9] Lam N N-S. Spatial Interpolation Methods: a Review. *American Cartographer*; 1983, **10**: 129–149.

[10] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM-algorithm. *Journal of the Royal Statistical Society*; 1997; **39** (1) :1–38.

[11] Banerjee S, Gelfand, A E, Bradley P C. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC; 2004.

[12] Aster R, Borchers B, Thurber C. *Parameter Estimation and Inverse Problem* (*International Geophysics Series*). Academic Press; 2005.

[13] Yoo E H and Kyriakidis P C. Area-to-point Kriging in spatial hedonic pricing models, *Journal of Geographical Systems*; 2009; **11**(4):381–406.

[14] Yoo E H and Kyriakidis P C. Area-to-point Kriging with inequality-type data, *Journal of Geographical Systems*; 2006; **8**(4):357–390.

[15] Mugglin A S, Carlin B P, Zhu L, Conlon E, Bayesian Areal Interpolation, Estimation, and Smoothing: An Inferential Aroach for Geographic Information Systems: Population Interpolation Over Incompatible Zones, *Journal of Agricultural, Biological, and Environmental Statistics*; 1999; **3**:117–130.